УДК 159.9.681.5

АНАЛИЗ СТРУКТУРЫ МНОГОМЕРНЫХ ДАННЫХ МЕТОДОМ ЛОКАЛЬНОЙ ГЕОМЕТРИИ

О.Г. Берестнева, Е.А. Муратова, А.Е. Янковская

Томский политехнический университет
E-mail: ean@rambler.ru

Рассматриваются один из подходов анализа структуры многомерных данных методом локальной геометрии. Обосновывается разработка интеллектуальных систем, способных адаптироваться к конкретным прикладным задачам, учитывать особенности исследуемых данных и строить вычислительный процесс в зависимости от полученных результатов.

Введение

Одним из методов получения наглядного визуального представления о логических закономерностях в структуре данных является метод локальной геометрии. В отличии от традиционных методов анализа многомерных данных, которые используют представление об общем пространстве признаков и об одинаковой мере сходства и различия, в методе локальной геометрии каждый объект рассматривается как самостоятельный классификатор, и для него строится собственное (локальное) пространство признаков, в котором определяется индивидуальная мера сходства и различия с другими объектами [1].

Использование метода локальной геометрии для обнаружения закономерностей в базах данных позволяет получать следующие преимущества, указанные в [1]: 1) достаточно простое построение IF...THEN правил в данных; 2) устойчивость закономерностей проверяется с помощью множества фальсификаторов; 3) выявляется структура логических закономерностей в данных; 4) достигаются минимальные ошибки при решении задач классификации, распознавания образов и прогнозирования.

Анализ геометрической структуры данных методом локальной геометрии не имеет готовых шаблонов и реализуется известными методами и алгоритмами, использующие геометрическое описание

данных. Например, для реализации этих методов можно воспользоваться математическими и графическими средствами статистических пакетов StatGraphics, Statistica, SPSS и др. Однако применение данных пакетов сопряжено с рядом трудностей: они англоязычны, требуют знаний статистических методов. Следует отметить и тот факт, что сложность структуры экспериментальных данных и разноплановость задач, например, при проведении медико-психологических исследований, существенно затрудняют применение вышеупомянутых пакетов анализа данных.

Поэтому разработка интеллектуальных систем, способных адаптироваться к конкретным прикладным задачам, учитывать особенности исследуемых данных и строить вычислительный процесс в зависимости от полученных результатов (например, разведочным анализом данных) является актуальной. Исследователю, работающему с такой интеллектуальной системой, становится доступным инструментарий, который позволит в интерактивном режиме изучить закономерности, имеющиеся в структуре исходных данных.

Разрабатываемая нами система интеллектуального анализа данных предполагает:

– самостоятельно получать новые знания об особенностях объектов исследования;
– учитывать локальные особенности (опора на прецедент) в каждой конкретной точке пространств решений;
– изучать структуру исходных многомерных данных с помощью отображения в двумерное пространство;
– исключать из дальнейшего исследования неинформативные признаки;
– выбирать методы обработки информации в процессе решения задачи;
– дать возможность исследователю самостоятельно определять последовательность этапов анализа диагностической информации на базе уже достигнутых результатов.

Такой процесс непрерывного взаимодействия исследователя со своим "интеллектуальным" помощником создает предпосылки для формирования своеобразного "гибридного" интеллекта, который обеспечивает эффективное использование достоинств объектов разной физической природы при взаимной компенсации их недостатков [2]. Кроме того, данный подход позволяет осуществлять контроль правильности ввода данных, построение решающих правил, формирование баз данных и знаний.

Основные подходы к анализу структуры многомерных данных

Реальные данные экспериментов в неструктурированных прикладных областях знаний, как правило:

1) не имеют сведений о законах и параметрах распределений,
2) для них ничего не известно о степени представительности выборки,
3) неоднородны и разнотипны,
4) имеют пробелы и ошибки, шумящие, неинформативные признаки,
5) отличаются высокой размерностью признакового пространства.

Поэтому применение точных методов представляется не целесообразным.

В целом, следует отметить, что сейчас достаточно развиты как параметрические, так и непараметрические методы анализа структуры данных, в том числе и для задач сокращения размерности и классификации [1–6]. Однако, разработка интеллектуальных систем, способных анализировать сложную структуру экспериментальных данных неструктурированных областей знаний, а также формировать и оптимизировать базу знаний до сих пор остается актуальной [2, 4].

Выбор математического аппарата и формализация знаний самым существенным образом зависят от проблемной области. Именно адекватность используемой математической модели реальному положению вещей, корректность представления знаний в рамках выбранной модели и эффективность применяемых математических преобразований определяют в конечном итоге оперативность и качество решений, которые будут приниматься по результатам распознавания произвольных объектов из данной проблемной области.

Интерактивность при выполнении задач анализа структуры многомерных данных связана с передачей исследователю ряда трудно формализуемых или технически сложно реализуемых операций. При этом мощность вычислительной техники расходуется как на выполнение отдельных этапов обработки (в соответствии с указаниями пользователя), так и на наглядное представление информации исследователю на этапах ее изучения, выработке решений по дальнейшей обработке и интерпретации результатов.

В пользу режима интерактивного взаимодействия приведем следующие факты.

Во-первых, исследователю заранее не известно, какая математическая модель наиболее подходит для решаемой задачи, и выбор одного из известных на сегодня стандартных пакетов прикладных программ обработки данных не располагает "всеобщим" методом построения адекватной всем задачам распознавания математической модели. Существует набор методов решения задач анализа данных, не связанных однозначно друг с другом.

Во-вторых, основная трудность проведения компьютерного анализа экспериментальной информации состоит в выборе способа описания исходных данных, методов их упорядочения и клас-

сификации, а также при необходимости получения оценок параметров распределения этих данных. Точность таких оценок, их устойчивость и прогностическая эффективность в значительной мере зависят от того, по отношению к какому множеству данных эти оценки получены. Поскольку данные, например, медицинских и психодиагностических исследований часто представляют собой относительно малые по величине и нерепрезентативные обучающие выборки, то применение параметрических методов статистики для анализа таких данных исключается, а использование известных методов группировки данных (методы таксономии, кластерный анализ, факторный анализ и др.) ограничивается.

В-третьих, известно, что человек имеет несомненные преимущества перед алгоритмами при распознавании некоторых видов структур – кластеров, однородных групп, когда размерность пространства описания не превышает трех. А, именно, алгоритмы автоматической классификации не могут правильно разделить исходное множество на группы при наличии точек соприкосновения и частичного перекрытия групп, при объединении в одном кластере нескольких удаленных друг от друга групп и при сложной форме кластеров.

В-четвертых, полученные экспериментальные данные всегда сопровождаются искажениями, вызванными шумами и помехами, которые легко "сбивают" алгоритмы автоматической классификации и идентификации. Эти искажения ограничивают также и возможности человека при самостоятельном анализе данных. В интерактивном режиме взаимодействия реализуются операции и средства для очищения информации от шума.

В данной работе структура многомерных данных анализируется с использованием методов, основанных на геометрическом представлении данных, в виде точечных скоплений в двумерном пространстве описаний. Для визуализации данных воспользуемся так называемыми distant-алгоритмами [2–4, 6, 9, 10], которые используют в качестве меры упорядочения значения взаимных расстояний между точками-образами.

Таким образом, задачи анализа многомерных данных могут быть сведены к трем основным:

– классификации исходных данных,
– выбору информативных признаков,
– идентификации неизвестных наблюдений.

Все эти задачи можно представить как варианты задачи группирования, которые в режиме интерактивного взаимодействия решаются с позиций принципа визуального группирования [1, 4, 6].

Если определить классификацию как объединение элементов выборки в подмножества с помощью того или иного правила (критерия), которая позволяет выявить схожие элементы, то она, по существу, совпадает с задачей упорядочения элементов, близких по значениям признаков.

В режиме интерактивного взаимодействия в соответствии с принципом визуального группирования основным критерием для объединения (или разделения) элементов выборки в одну группу является близкое расположение (или удаление) точек-образов этих элементов друг к другу. Формирование групп, определение конфигурации области и количества включаемых в группу точек целиком зависит от мнения исследователя.

Задача выбора информативных признаков сводится к задаче группирования (если выбор информативных признаков понимать как поиск группы признаков, сходных по своему проявлению на элементах выборки) с последующей оценкой отдельных признаков или их групп на сохранность структуры упорядочения, полученной на полном перечне признаков.

Исключив из описания выборки признак или группу признаков и применив алгоритм визуализации, можно получить отображение структуры выборки в виде точечного скопления. Сравнив полученные изображения, исследователь принимает решение об информативности группы признаков, т.е. о сохранении или исключении этих признаков из описания выборки.

При решении задачи распознавания предполагается отнесение неизвестного наблюдения к одному из известных классов, на которые разбита исходная выборка. Если в качестве основного принципа для построения стратегии решения этой задачи взять принцип визуального группирования, то и эту задачу можно свести к задаче группирования. Ее можно определить как упорядочение совместной выборки, включающей исходную выборку и неизвестное наблюдение. Полученное изображение точечного скопления предъявляется исследователю, а он решает, к какой из групп следует отнести неизвестный элемент. Естественно, что это решение основано на оценке близости (удаленности) расположения точки-образа вновь введенного элемента к точечному скоплению одной из групп (в общем случае возможно и другое решение – объект не может быть отнесен ни к одной из групп).

Анализ структуры многомерных данных методом локальной геометрии

Анализ структуры многомерных данных с использованием метода локальной геометрии базируется на комбинированном применении методов линейной алгебры и интерактивной графики.

Постановка задачи не нова и содержится в [1]. Схема анализа структуры многомерных данных, позволяющая осуществить поиск логических закономерностей в локальном пространстве признаков заключается в следующем.

На первом этапе с целью унификации признакового пространства осуществляется преобразование исходных признаков в бинарные или k-значные признаки посредством модифицированного

нами алгоритма адаптивного кодирования признакового пространства. Преобразование основано на предположении о том, что признаки, можно дискретизировать таким образом, чтобы отношение относительных частот встречаемости объектов обучающих выборок двух классов в выделенных интервалах могло быть аппроксимировано одноэкстремальной или монотонной функцией. Это позволит, в зависимости от типа признака, каждому выделенному интервалу присвоить кодовое число, или в случае преобразования в бинарные признаки, использовать выделенные интервалы в качестве самостоятельных признаков [7, 8].

На втором этапе с целью анализа структуры многомерных данных предлагается реализовать подход, предложенный в [1].

Для равномерного распределения объектов исследования в исходном пространстве признаков вводятся в альтернативные классы "шумящие" объекты, представляющие собой множество фальсификаторов, "столкновение" с которыми способствует лучшему проявлению устойчивых логических закономерностей в данных.

Для определения наиболее перспективного объекта, относительно которого строится локальное пространство признаков, данные отображаются на плоскости двух первых главных компонент. Выбор последующих центральных объектов ведется в соответствии с целью исследования, например, в качестве цели может быть выбрано изучение объектов, расположенных за границами выделяющихся точечных скоплений. В ходе исследования цели могут корректироваться с учетом обстоятельств текущего анализа.

После центрирования данных относительно выбранного объекта применяют один из методов определения локальных взвешенных метрик. В качестве метода определения локальной взвешенной метрики, например, могут быть использованы методы конструирования линейных диагностических решающих правил, факторный анализ, методы эволюционного моделирования, а также ряд других методов [2–4, 9, 10]. Оценивание построенных локальных взвешенных метрик производится по следующему критерию [1]

$$J = \frac{\sum d(x, x_j)}{\sum d^{'}(x, x_k)} = \min,$$

где $\Sigma d(x,x_l)$ – суммарное расстояние от объекта до объектов своего класса, а $\Sigma d'(x,x_k)$ – суммарное расстояние объекта x до объектов других классов; либо посредством визуального анализа гистограмм распределения расстояний от объектов обучающей выборки до исследуемого объекта.

Средствами интерактивной графики, осуществляемой после визуализации данных, из анализа исключаются наиболее удаленные от нулевой отметки новой оси объекты, признаки с отрицательными весовыми коэффициентами (для сохранения метрических соотношений), объекты, имеющие равные расстояния с объектами других классов.

После того, как построены локальные взвешенные метрики (линейные классификаторы), необходимо изучить взаимодействие данных классификаторов. Для этих целей можно воспользоваться методами построения коллективных решающих правил. После проверки нарушений метрических отношений в матрице расстояний, которые могут возникнуть из-за различия пространств описания локальных классификаторов, исследование структуры матрицы расстояний может производиться методами и алгоритмами, использующими геометрическое описание данных.

При выборе решающего логического правила из системы диагностических правил приоритет отдается правилу, обладающему наибольшей эффективностью при распознавании исследуемых объектов.

Анализ структуры многомерных данных с применением локальной геометрии позволяет оставить в описании только то, что действительно важно для отражения сходства и различия с другими объектами. Это обеспечивает каждому объекту, как представителю своего класса, максимально возможную "сферу действия", чего нельзя достигнуть при построении общего пространства признаков и использовании одинаковой метрики для всех объектов.

Структура интерактивной системы интеллектуального анализа многомерных данных

Разрабатываемая интерактивная система интеллектуального анализа данных программно реализуется с использованием инструментального средства Borland Delphi 5.0 в среде Windows 95/98/2000/NT. Структура системы приведена на рисунке.

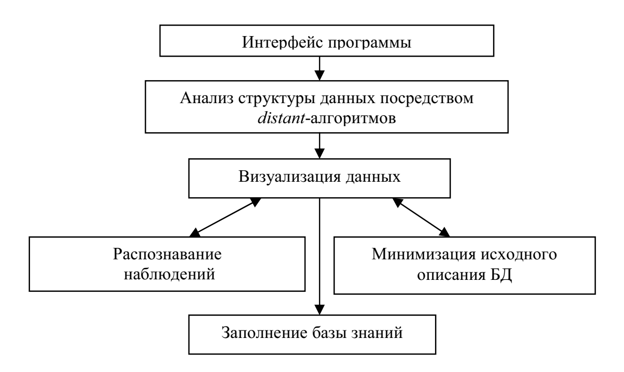


Рисунок. *Структура системы интеллектуального анализа данных*

Заключение

Предложенный метод анализа структуры многомерных данных позволит: сделать более понятными критерии и принципы построения правил вхождения объектов в определенные классы эквивалентности; минимизировать ошибки при принятии решений; находить скрытые в больших объемах данных закономерности в структуре данных, зачастую не формулируемые экспертом, и пополнять ими базу знаний системы; получать результирующую интерпретацию и аргументированные ответы на вопросы, какие закономерности лежат в основе диагностического заключения.

В дальнейшем предполагается сравнение данного подхода выявления закономерностей с подходом, предложенным в работе [11].

Работа поддержана частично грантами РФФИ (проект № 01-01-01050, № 01-01-00772, № 03-06-80128) и грантами РГНФ (проекты № 01-06-00084а, № 02-06-00086а)

СПИСОК ЛИТЕРАТУРЫ

1. Дюк В.А. Компьютерная психодиагностика. — СПб.: Изд-во "Братство", 1994. — 364 с.
2. Загоруйко Н.Г. Прикладные методы анализа данных и знаний. — Новосибирск: Изд-во Ин-та математики, 1999. — 270 с.
3. Журавлев Ю.И. Об алгебраических методах в задачах распознавания и классификации // Распознавание, классификация, прогнозирование. Математические методы и их применение. — М.: Наука, 1989. — Вып. 1. — С. 9—16.
4. Прикладная статистика: Классификация и снижение размерности: справ. издание / Под ред. С.А. Айвазяна. — М.: Финансы и статистика, 1989. — 608 с.
5. Левит В.Е., Переверзев-Орлов В.С. Структура и поле данных при распознавании образов. — М.: Наука, 1984. — 124 с.
6. Попечителев Е.П., Романов С.В. Анализ числовых таблиц в биотехнических системах обработки экспериментальных данных. — Л.: Наука, 1985. — 148 с.
7. Таранова Н.Н. Метод адаптивного кодирования признаков // Динамика систем. Межвуз. тематич. сб. научн. тр. / Под ред. Ю.И. Неймарка: Нижний Новгород: Нижегор. гос. ун-т, 1995. — С. 54—70.
8. Янковская А.Е., Муратова Е.А., Берестнева О.Г. Унификация разнотипных данных в интеллектуальных распознающих системах // Знание-Диалог-Решение (KDS-2001). Труды Международной научно-практ. конф. Том 2. — СПб.: Лань, 2001. — С. 661—668.
9. Журавлев Ю.И., Камилов М.М., Туляганов Ш.Е. Алгоритмы вычисления оценок и их применение. — Ташкент: ФАН, 1974. — 38 с.
10. Yankovskaya A.E. Minimization of Orthogonal Disjunctive Normal Forms of Boolean Function to be Used as a Basis for Similarity and Difference Coefficients in Pattern Recognition Problems // Pattern Recognition and Image Analysis. — 1996. — Vol. 6. — No 1. — P. 60—61.
11. Янковская А.Е. Логические тесты и средства когнитивной графики в интеллектуальной системе // Новые информационные технологии в исследовании дискретных структур: Докл. 3-ей Всеросс. конф. с международным участием. — Томск: Изд-во СО РАН, 2000. — С. 163—168.

УДК 539.3